Model-based clustering with sparse matrix mixture models

13th Scientific Meeting - Classification and Data Analysis Group

Alessandro Casa

Joint work with: Andrea Cappozzo & Michael Fop

School of Mathematics and Statistics
University College Dublin

alessandro.casa@ucd.ie

11th September 2021

Framework

- Matrix-variate (three-way) are increasingly widespread
  $\Longrightarrow$ multiple variables are measured on a set of of units
  in different occasions. Examples are
  - Longitudinal data with multiple features
  - Spatio-temporal and spatial multivariate data
  - Multi-attribute ratings by multiple experts
- Rather complex structure, need to account for the three layers

Idea

model-based clustering strategies might help
to uncover interesting patterns in the data

$\Downarrow$

Standard Gaussian Mixture Model are not
appropriate in this framework

Matrix Gaussian mixture model

- Let $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ be a set of n matrices with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$
- Matrix Gaussian mixture model (MGMM) represents the GMM extension for three-way data and is expressed as

$$f(\mathbf{X}_i; \Theta) = \sum_{k=1}^{K} \tau_k \phi_{p \times q}(\mathbf{X}_i; \mathbf{M}_k, \Omega_k, \Gamma_k)$$

  - $\phi_{p \times q}(\cdot, \mathbf{M}_k, \Omega_k, \Gamma_k)$, $p \times q$ matrix normal distribution
  - τ_k's, mixing proportions $\tau_k > 0, k = 1, \dots, K, \sum_k \tau_k = 1$
  - $\mathbf{M}_k$, k-th component mean matrix
  - Ω_k and Γ_k rows and columns precision matrices with dimensions $p \times p$ and $q \times q$ respectively
- Need to estimate $\Theta = \{\tau_k, \mathbf{M}_k, \Omega_k, \Gamma_k\}_{k=1}^{K}$

Overparameterization in action

- Limitation: $|\Theta|$ scales quadratically with p and q

$$\Downarrow$$

  - Dramatic overparameterization even with moderate dimensions
  - Difficult interpretation of the relations among variables/occasions across the clusters

- Proposed solutions introduces a rigid way to induce parsimony $\Longrightarrow$ association structures constant across groups

Our assumption

the matrices in Θ possess some cluster-dependent degrees of sparsity

Sparse matrix-variate mixture model

- We maximize a penalized log-likelihood defined as

$$\ell(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X_i}; \mathbf{M}_k, \Omega_k, \Gamma_k) - p_{\lambda_1,\lambda_2,\lambda_3}(\Theta)$$

where $p_{\lambda_1,\lambda_2,\lambda_3}(\Theta)$ is equal to

$$\sum_{k=1}^{K} \lambda_1 ||\mathbf{P}_1 * \mathbf{M}_k||_1 + \sum_{k=1}^{K} \lambda_2 ||\mathbf{P}_2 * \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||\mathbf{P}_3 * \Gamma_k||_1$$

  - $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ matrices with non-negative entries
  - $\lambda_1, \lambda_2, \lambda_3$ penalty coefficients
  - $||A||_1 = \sum_{jh} |A_{jh}|$

- Advantages
  - Less parameters + easier interpretation
  - Irrelevant variables detection
  - Cluster-wise conditional dependence patterns

Parameter estimation

- EM-algorithm to maximize a penalized complete log-likelihood

$$\ell_c(\Theta; \mathbf{X}) \propto \sum_{i,k} z_{ik} \Big[\log \tau_k + \frac{q}{2} \log |\Omega_k| + \frac{p}{2} \log |\Gamma_k| + \\ -\frac{1}{2} \mathrm{tr} \left\{ \Omega_k (\mathbf{X}_i - \mathbf{M}_k) \Gamma_k (\mathbf{X}_i - \mathbf{M}_k)^\intercal \right\} \Big] - p_{\lambda_1, \lambda_2, \lambda_3}(\Theta)$$

- E-step $\Rightarrow$ standard updating formula
- M-step $\Rightarrow$ partial optimization strategy
  - τ_k: standard update
  - $\mathbf{M}_k$: sparsely estimated via cell-wise coordinate ascent (matrix-variate extension of Th.1 in Pan et al., 2009)
  - Ω_k and Γ_k: sparsely estimated via suitable modification of the coordinate descent graphical LASSO algorithm

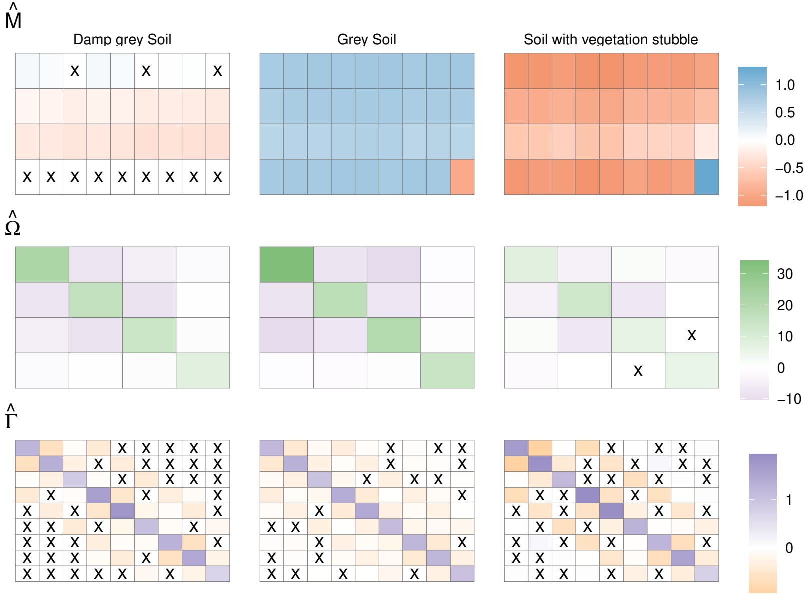
Some results - Satellite image dataset

- $n = 845$ satellite images belonging to $K = 3$ classes (grey soil, damp grey soil, soil with vegetation stubble)
- Images represented by $q = 9$ pixels and recorded $p = 4$ times
- Data can be represented as a collection of 845 matrices having dimensions 4×9

	Sparse MGMM	**PMGMM**	**Mclust**
Adjusted Rand Index	0.7883	0.7772	0.3841
# of free parameters	218	275	850

- Good recovering of the clustering structure with more decreased number of estimated parameters

Some results - Satellite image dataset



Concluding remarks and future directions

- We propose a penalized estimation strategy for MGMM
- Reduction of the number of parameters to estimate, flexible way to induce parsimony, enhanced interpretation
- Chance to resort to *Mix* & *Match* approaches

- Open problems $\Rightarrow$ we need to select $\lambda_1, \lambda_2, \lambda_3$ and K
  - Exhaustive search and approaches based on cross-validation are computationally unfeasible
  - Conditional search, E-MS algorithm, Genetic algorithm?
  - Every suggestion is more than welcome

Some references

- Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3): 432–441.
- Sarkar, S., Zhu, X., Melnykov, V. & Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Computational Statistics and Data Analysis*, 142:106822.
- Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4): 511-522.
- Zhou, H., Pan, W. & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3: 1732-1496.